CEIS

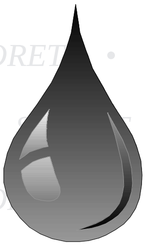
center for environmental information and statistics

Center for
Environmental
Information and
Statistics

US Environmental
Protection Agency

Major Findings from the CEIS Review of EPA'S

STORET DATABASE



July 13, 1999

Major Findings from the CEIS Review of EPA's STORET Database

1. INTRODUCTION

The new, modernized **STORET** is EPA's principal repository for marine and freshwater ambient water quality and biological monitoring information. It combines the functions of the original STORET with that of the Biological Information System (BIOS) and the Ocean Data Evaluation System (ODES). These systems have served as the Agency's primary sources of point and non-point source ambient water quality and biological monitoring data. Their analytical tools supported a wide range of EPA water quality and ecosystem health assessment activities. Together, these systems contain over 250 million parametric observations collected primarily by State agencies from over 700,000 sampling stations nationwide, representing an investment of over $2.2 billion.

The original STORET was developed in the 1960's and served as the nation's primary water quality storage system for 33 years. Program requirements and technology have outpaced the original STORET. It was unable to track monitoring information such as how, why, and by whom water samples were taken, thus making the quality of the data questionable and limiting its usefulness. The system also suffered from year 2000 problems that would have been prohibitively expensive to correct. Data from the legacy system will be reviewed to ensure that it meets current quality requirements before migrating it to the modernized system. All pre-1999 data, irrespective of whether it meets current requirements, will be maintained in the Legacy Data Center (LDC) and made available to the public as long as it is technically feasible and cost effective.

The first version of the new STORET software to collect data was released in September 1998, and STORET Version 1.1 was released in March 1999. The new system addresses the problems with the legacy version. It is easier to use, supports the storage of quality assurance and quality control information, is flexible enough to handle the changing needs of its users, and provides a wide range of standard output formats. STORET currently also supports the Geographic Information System environment. Future updates and revisions are planned to maintain a dynamic system. While the software to input the data in the system has been distributed to data providers, the new system is not yet populated with the data. It is expected to be operational by the end of 1999.

STORET currently better meets the emerging data and information needs associated with watershed-level environmental protection. It promotes data sharing and meets spatial assessment requirements for successful local watershed protection programs. The water monitoring community will have access to information that accurately reflects the current status while illustrating future monitoring and assessment paths. Decision makers can use STORET both to plan and to evaluate the effectiveness of pollution prevention and abatement programs.

The modernized STORET was designed with several purposes in mind: to investigate water quality where there have been reports of poor water quality or where an event affecting water quality has occurred; to provide the data needed to make a continuing broad assessment of water quality within a geographic area for which the data provider holds responsibility for protecting and improving water quality; and to support systematic assessments of data providers' watersheds.

This review is limited because a number of questions cannot be answered until the system has been in use for an extended period of time. EPA's Office of Water will continue to administer and maintain STORET.

2. SUMMARY ANSWERS TO REVIEW QUESTIONS

2.1. What does the database cover?

STORET is EPA's principal repository for marine and freshwater ambient water quality and biological monitoring information, and includes data on the chemical composition of water, biological community information, sediment toxicity information, fish tissue analysis, and aquatic habitat evaluations.

It is impossible to determine the comprehensiveness or physical size of the modernized STORET database as it is not currently populated. A rough estimate can be drawn from the size of the legacy STORET database, which has a total of about 700,000 monitoring stations, some of which are more continuously monitored than others.

STORET has no reporting requirements except State-generated 305(a) and 305(b) reports required by the Clean Water Act, but it is used extensively in other applications by States as well as in the Total Maximum Daily Load (TMDL) program. All non-305 data placement within STORET is done voluntarily. Because of this, STORET's spatial comprehensiveness is variable and dependent upon the contributions of State and Federal agencies, as well as others in the watershed monitoring community such as tribes, local governments, academic groups, watershed organizations, and citizen volunteers.

Legacy data from the original STORET will be examined and migrated into the modernized system if it meets the current quality requirements. All legacy data will also be archived for retrieval in the Legacy Data Center (LDC).

2.2. Can the database be used for spatial analysis?

Geographic analysis is possible using latitude and longitude information.

2.3. Can the database be used for temporal analysis?

Temporal analysis is possible. Data from the Legacy Data Center (LDC) are hampered by non-standardized data collection and reporting regimens. Data in the new system (both newly collected and migrated from the old system) are expected to be more standardized and amenable to temporal analysis.

2.4. How consistent are the variables over space and time?

There is no formalized consistency of variables over space and time. Consistency is a function of data providers' collection programs, and these programs vary widely in purpose and scope.

2.5. Can data from STORET be linked with information from other databases?

STORET contains latitude and longitude data that can be used to link with information from other databases. STORET is also Environmental Data Registry (EDR) compliant, and efforts are underway to link data more effectively with the US Geological Survey (USGS) data.

2.6. How accurate are the data in STORET?

With few exceptions, data accuracy is solely the responsibility of the data providers. EPA does not undertake quality audits for STORET; therefore, it is extremely difficult to assess the accuracy of the data.

2.7. What are the limitations of STORET?

STORET will likely be limited by its flexibility, by data provider ownership, and by its largely self-policing nature. There are extremely limited controls placed on data providers, largely to encourage them to provide data to the system. This lack of centralized control may result in non-standardized reporting, thus limiting the utility of the database.

2.8. How can I get information on STORET?

Only the legacy data are currently available from the EPA. Legacy and current data from STORET should be available in 1999 via the internet. Summarized STORET information is available in the Surf your Watershed and Index of Watershed Indicators databases via EPA's web site. Data from STORET may be obtained from:

STORET User Support
Environmental Protection Agency
401 M Street, SW, Mail Code 4503F
Washington, DC 20460
Phone: (800) 424-9067
email: STORET@epa.gov

2.9. Is there documentation on STORET?

There are a number of user and system documents available from the Office of Wetlands, Oceans, and Watersheds.

3. DETAILED ANSWERS TO REVIEW QUESTIONS

3.1. What does the database cover?

STORET is maintained by the EPA for the storage and retrieval of chemical, physical, and biological data pertaining to the quality of waterways within and contiguous to the United States. The legacy database has served primarily as a repository of parametric water quality data since the 1960's, including chemical composition of water, biological community information, sediment toxicity information, fish tissue analysis, and aquatic habitat evaluations. The database has also evolved into a family of systems containing geographical, political, and descriptive information about sites where data have been collected; counts and descriptions of living organisms found at these sites; and stream flow data as obtained from the US Geological Survey (USGS). STORET continues to be EPA's principal repository for marine and freshwater ambient water quality and biological monitoring information.

The first version of the modernized STORET was released in September 1998, version 1.1 was released in March 1999, and subsequent revisions and updates are planned. Information on the comprehensiveness or physical size of the modernized STORET database is not available as it is not yet populated and the process of migrating data from the old to new system is not complete. A rough estimate can be drawn from the size of the legacy STORET database, which contains data for a total of 700,000 monitoring stations, with some stations monitoring and reporting more continuously than others.

STORET planners expect that some, but not all, of the data entered into the legacy STORET will migrate to the modernized STORET. Because the data in the legacy STORET were often not of a documented quality, the incomplete data cannot be entered directly into the new system, which has much more stringent data quality requirements. The data will therefore be migrated to the new STORET system in a two-step process. First, all old data will be moved into the Legacy Data Center (LDC). The LDC is a read-only archive that will provide a stable platform for this data, permit a more orderly and systematic approach to data migration, and solve technical problems with 1999 and 2000 dates in the legacy software. EPA will move old STORET data to the LDC by the fall of 1999. It will then be left to the owners, or original providers, of the data to see what data can be migrated from LDC to the current STORET based on required data documentation outlined by EPA. Some old data of undocumentable quality will never be migrated to the new STORET. Legacy data will be frozen and made available to the public through the internet.

Who Must Report?

The only information required to be submitted to STORET are State-generated 305(a) and 305(b) reports, which are required every two years by the Clean Water Act. In addition, States are strongly encouraged to participate in STORET by contributing watershed and waterbody data files. Beyond that, there are no reporting requirements, and the placement of all non-305 data within STORET is done voluntarily. Because of this, STORET's spatial comprehensiveness is variable and dependent upon the contributions of State water pollution control agencies and the Federal agencies that have provided data to legacy STORET, as well as others in the watershed monitoring community such as tribes, local governments, academic groups, watershed organizations, and citizen volunteers.

How are data reported?

Data are reported by data providers through self-entry into the new STORET database. The database incorporates a user-friendly data entry interface with a series of pull-down reference tables, or "pick lists," embedded in the database. These pick lists will require information to be entered item by item into the following fields: water quality characteristic; method used to obtain the measured value of that characteristic; the specific analytical instrument used to carry out that method; measurement units associated with the numerical value to be reported; and so forth. The sequence of pick lists will continue as appropriate for each context until the characteristic whose value is to be reported and the manner in which that characteristic was obtained is sufficiently well-specified, and two different observations contained in STORET having the same set of picked elements will in fact be comparable to each other. There is also a batch input capability for the new system.

Data Elements

STORET is now divided into two major sections: Data Maintenance and System Administration. Data elements described and recorded in the Data Maintenance section include: organizations; projects; sampling stations; trips; station visits; and monitoring results. The System Administration section houses the support functions necessary for the operation and maintenance of the database itself. Data entered into the Data Maintenance section are organized into five categories. Specific information is required under each of these categories in order for STORET to receive data.

Organization — This includes information on the group or entity responsible for the data set, either for collecting and otherwise generating the data or sponsoring the activity for which the data set was created. Organizations have the ability to document sampling results, as well as the entire process leading to the results, including how and with what equipment the sample was collected, how the sample was preserved and transported, and what sample preparation and analytical methods were employed by the laboratory to generate the results.

Projects — This specifies the activity during and for which the data set was created. An organization may have an unlimited number of projects. Project descriptions must be linked back to one or more stations before field results may be stored.

Stations — This category identifies and describes the physical location at which monitoring occurs. An organization may have an unlimited number of stations. Description of a station or study area must include a latitude/

longitude location and a State or State-county code or name. Additional descriptive information may also be recorded. Stations must be linked back to one or more projects if data are to be stored.

Trips, Visits, and Samples — This includes water quality sampling, observation, and measurement activities that occur at these sites, as well as comprehensive descriptors of the event during which samples were collected. Trips and station visits are the high level activities conducted in support of a project, and through which data collected in the field are associated with specific stations and dates. Trip descriptions include names of projects to be supported and stations to be visited. Station visits include actual field measurements and observations about a site and sample collection. Samples are described according to the medium sampled and the intent for which they were collected. Standard procedures followed in the collection of samples, including the handling and transport of samples, preparation done prior to lab analysis, lab methods, and equipment used are also fully described.

Results — This category includes findings of the sampling event, measurements, and field activities. Results can be entered in three ways: as a value for numeric quantities or measurements; as a selection from a “choice list” for certain physical characteristics provided by STORET; or as free text entries for general observations.

3.2. Can the database be used for spatial analysis?

STORET’s geographic universe of concern is determined independently by each data provision agency that enters data into the system. There is no universe of concern that is common to all such data provider agencies; however, each monitoring station is described by a specific combination of latitude and longitude and by a State or State and county code. Therefore, geographic analysis is possible, though of questionable utility given the non-standardized nature of data collection, treatment, and reporting activities among data providers for the legacy data.

3.3. Can the database be used for temporal analysis?

The function of STORET is dictated by its data, which are dependent upon its providers. There is no standard for either frequency of collection or reporting of temporal data. Sample collection frequency varies considerably from agency to agency and in many cases from station to station for each individual reporting group. Data collection frequencies may be monthly, yearly, or even daily for intensive surveys. Although temporal analysis is currently technically possible, the practicality of such analyses is dependent largely on the consistency of reporting.

3.4. How consistent are the variables over space and time?

Because of the completely decentralized nature of STORET, there is no guidance regarding the consistency of variables over time or space. There are, however, some required and consistent ways of identifying each characteristic for data being monitored and entered into the database. This provides some consistency in describing data, but not in maintaining the consistency of the database itself. Maintaining consistency is left to the discretion of the data owners and providers. It is ultimately up to each individual agency to determine how and when to collect the data. There continues to be great variation among these agencies as to the design and conduct of their current data collection programs.

3.5. Can data from STORET be linked with information from other databases?

An effort has been made to facilitate the use of STORET by EPA's Drinking Water Program. The Drinking Water Program has agreed to use the new STORET structure and characteristics as the starting point for developing its new National Occurrences Database in which concentrations of monitored drinking water contaminants will be stored. In addition, there are latitude and longitude data that can be used to link with information from other databases.

3.6. How accurate are the data in STORET?

Accuracy of the data is largely determined by the data providers. Currently, STORET staff have editing privileges only in those cases where there are clear limits on the feasible range of values of a water quality characteristic (e.g., pH levels). All other data quality audits are the responsibility and province of data providers. STORET's pick-list data entry system will incorporate pull-down tables that require data providers to select the type of data analysis procedures acceptable to EPA for any given water quality characteristic. However, compliance with these procedures will be entirely by self-policing. Therefore, the true accuracy of the data cannot be known.

3.7. What are the limitations of STORET?

STORET will likely be limited by its flexibility, by data provider ownership, and by its largely self-policing nature. The system is designed as a central repository for water quality data, and is not designed to impose restrictions and regulations on State and local water quality data collection activities. Rather, it permits users to access a vast array of data collected on watersheds throughout the country. While STORET now requires quality assurance information that was not required by the legacy system, such as when, where, how, why, and by whom data samples were collected, it is still possible that users will assume that the data are more reliable than they actually are.

The basis of STORET is data-provider ownership. EPA encourages the use of STORET by the supplying agencies as their use helps police the system. This system has both advantages and disadvantages; prime among the disadvantages is the reliance on responsible provision of data. Data accuracy is almost the sole responsibility of data providers. Should erroneous data be found in the system, and the original data provider refuses to change the data and does not grant EPA permission to change the data, the data will remain in the system. Also, data providers will be able to delete any of their data at any time. This degree of ownership facilitates State and local agency buy-in. On the other hand, it also provides for the possibility of inconsistent data.

STORET's data providers are expected to be self-policing. That is, the onus for reporting accurate, well-collected data rests with the reporter. While there is no real benefit to reporting bad data, the system has sufficient flexibility to allow such abuses to occur.

3.8. How can I get information on STORET?

To receive a copy of the software for the new STORET, access data, or obtain additional information about STORET contact:

STORET User Support
Environmental Protection Agency
401 M Street, SW, Mail Code 4503F
Washington, DC 20460
Phone: (800) 424-9067
email: STORET@epa.gov

Data from STORET will be available in 1999 via the internet. In addition, summarized legacy STORET information is available in the Surf your Watershed (http://www.epa.gov/surf/) and Index of Watershed Indicators (http://www.epa.gov/surf/iwi/) databases via EPA's web site.

Currently, some background information on STORET is available electronically at:

http://www.epa.gov/OWOW/STORET/modern/anitest.html

3.9. Is there documentation on STORET?

There is documentation on monitoring guidance, on 305(a) reporting, and on data collection available from the Office of Wetlands, Oceans, and Watersheds. EPA also provides training on the design and use of STORET, including data editing and summarization/estimation procedures.